ISSN 2007-9737

La predicción lineal aplicada al reconocimiento distribuido del habla en redes IP

Elieser E. Gallego[1], Angel Correa[1], Alexei Blanco[1], Jessica Cuador[2], Raúl P. Álvarez[3]

[1] Universidad de Pinar del Río Hermanos Saíz Montes de Oca,
Departamento de Telecomunicaciones y Electrónica,
Cuba

[2] Universidad de Piar del Río Hermanos Saíz Montes de Oca,
Departamento de Matemática,
Cuba

[3] Universidad de Piar del Río Hermanos Saíz Montes de Oca,
Departamento de Forestal y Agronomía,
Cuba

{elieser, angel, alexei, jcuador, ruly}@upr.edu.cu

Resumen. Este trabajo presenta resultados en la aplicación de técnicas de compresión al reconocimiento del habla en redes IP. Por lo que las tres áreas fundamentales del conocimiento que están involucradas son: la compresión de voz, la simulación de canales de transmisión y el reconocimiento del habla. Dichas áreas se integran para lograr una aplicación desarrollada en MATLAB, capaz de realizar un reconocimiento del habla de forma remota mediante la transmisión de voz comprimida, con la peculiaridad de que se considera la probabilidad de ocurrencia de pérdidas de información en los canales de transmisión, de forma tal que queda demostrada la robustez del sistema propuesto.

Palabras clave. Reconocimiento del habla, compresión de voz.

The Lineal Prediction Applied to Distributed Speech Recognition in IP Nets

Abstract. This work presents results in the application of speech compression to speech recognition technics in IP nets. That's why the three fundamental areas of the knowledge that are involved are: the voice compression, the transmission channels simulation and the speech recognition. This areas are integrated to achieve an application developed in MATLAB, able to carry out a remote speech recognition by means of the transmission of compressed voice, with the peculiarity that it is considered the probability of occurrence of losses of information in the transmission channels, in such way that the robustness of the proposed system is demonstrated.

Keywords. Speech recognition, voice compression.

1. Introducción

Las técnicas de reconocimiento automático del habla tienen un constante desarrollo y evolución, además de una gran cantidad de posibles aplicaciones.

En ese sentido, la correcta interpretación del habla en la parte del receptor, tiene una gran importancia, sobre todo cuando de un correcto

Fig. 1. Aplicación del reconocimiento por voz

reconocimiento dependa la realización de acciones que requieran de determinado grado de precisión, seguridad, etc.

Las telecomunicaciones juegan un papel protagónico para el reconocimiento remoto, ya que permiten la transmisión de la voz cuyo significado audible es precisamente la información para el reconocimiento. Para lograrlo, es vital el aprovechamiento de las características espectrales de la señal de voz; lo cual re refleja directamente en el comportamiento de los coeficientes de predicción lineal LP (*Lineal Prediction Coefficients*), como se verá más adelante.

En cuanto a la transmisión de voz, en telecomunicaciones es una generalidad el empleo de codificadores (códecs o compresores) de voz. Esto se debe a la necesidad de realizar un uso cada vez más eficiente de los anchos de banda de los sistemas de telefonía IP, o sea, aquellos que emplean el Protocolo de Internet para el transporte de la señal de voz a través de las redes de datos.

Dentro de los códecs más conocidos se encuentran: G.711 (PCM, *Pulse Code Modulation*), G.728 (LD-CELP, *Low Delay Code Excited Lineal Prediction*) G.729 (CS-ACEPL, *Conjugated Structure – Algebraic Code Excited Lineal Prediction*), ILBC (*Internet Low Bit Rate*), entre otros. Los que mayor éxito tienen en cuanto a la reducción de la tasa binaria de transmisión son aquellos que emplean predicción lineal, como la norma FS1015 (LPC-10, *Lineal Prediction Coefficients*) que logra realizar una asombrosa compresión a 2.4kb/s.

El método de la predicción lineal se basa en la determinación de la muestra actual mediante una combinación lineal de muestras anteriores según la ecuación (1):

$$\tilde{S}(n) = \sum_{k=1}^{p} \propto_k * s\,(n-k), \qquad (1)$$

$$s(n) - \tilde{S}(n) = e(n), \qquad (2)$$

donde $\propto_k$ son los coeficientes de predicción lineal, $s(n)$ es la señal original y $\tilde{S}$ es la señal predicha [1].

El problema de la compresión empleando predicción lineal, es la determinación del conjunto de coeficientes $\propto_k$, para los cuales se minimiza el error de predicción según (2). En dependencia de cuál sea el códec empleado, será de un modo u otro el tratamiento de estos coeficientes [2].

A la hora de realizar reconocimiento remoto de voz, es importante considerar el tipo de compresión que se emplea, con el objetivo de garantizar la compatibilidad de las posibles aplicaciones de reconocimiento que puedan implementarse.

En este trabajo se presenta una propuesta para la realización del reconocimiento remoto del habla en redes IP, previa compresión empleando códecs cuyo algoritmo emplea la predicción lineal, y considerando la probabilidad de ocurrencia de pérdidas de paquetes en el canal de transmisión.

2. Materiales y métodos

En esta sección se abordan algunas técnicas, algoritmos y métodos a partir de los cuales se ha desarrollado esta investigación. Dentro de los que se encuentran: los códec empleados, los métodos de simulación y las técnicas de reconocimiento automático del habla.

– Los códecs empleados:

Al pretenderse realizar reconocimiento automático del habla sobre la voz comprimida por algún códec que emplee predicción lineal, se han elegido dos de los más representativos: el LPC-10 correspondiente a la norma norteamericana FS1015, y el CS-ACELP correspondiente al G.729. Esta elección se debe a que, el primero emplea la predicción lineal y transmite sus coeficientes sin otras modificaciones empleando un total de 54 bits para representarlos en la estructura de trama [3]; y el segundo una vez que determina los coeficientes los cuantifica en pares del espectro lineal (LSP, *Line Spectrum Pairs*) antes de enviarlos [4].

A continuación se describen los procesos que tienen lugar en estos códecs:

LPC-10: es un vocoder con una tasa binaria de 2,4kbps. Se corresponde con la norma

norteamericana FS-1015 [5]. Determina 10 coeficientes de predicción lineal para cada trama de 20mseg: calculando primero los coeficientes de auto correlación a partir de muestras de voz, ecuación (3), luego resuelve una matriz de Toeplitz que devuelve los 10 coeficientes de predicción, ecuación (4):

$$R_1(m) = \frac{1}{N}\sum_{n=0}^{N-1-m} X_1(n) * X_1(n+m) \qquad (3)$$

$$0 < m < p \ ,$$

$$\begin{bmatrix} r(0) & \cdots & r(p-1) \\ \vdots & \ddots & \vdots \\ r(p-1) & \cdots & r(0) \end{bmatrix}\begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} r(1) \\ \vdots \\ r(p) \end{bmatrix}, \qquad (4)$$

donde *r(0 p)* son los coeficientes de auto correlación de una trama X_n son las muestras de voz, *N* es el tamaño del segmento de muestras (160) y *p* es el orden de predicción, o sea, 10 [6].

G.729: es uno de los códec más usados en servicios de telefonía, por el positivo balance logrado entre la calidad perceptual y la baja tasa binaria.

Al igual que LPC-10, este códec emplea predicción lineal, pero esa información la cuantifica vectorialmente en pares de espectro lineal antes de enviarla.

Los coeficientes LSP q_i se cuantifican mediante la representación ω_i de LSF en el dominio normalizado de la frecuencia [*0, π*], o sea:

$$\omega_i = arcos(q_i) \ \ i = 1, \dots, 1. \qquad . \qquad (5)$$

Se aplica una predicción de media móvil (MA) de cuarto orden conmutada para predecir los coeficientes LSF de la trama en curso [3]. La diferencia entre los coeficientes calculados y su predicción se cuantifica mediante un cuantificador vectorial de dos fases. La primera fase es una cuantificación vectorial (VQ) de 10 dimensiones que utiliza una tabla de códigos L1 de 128 entradas (7 bits). La segunda fase es una VQ de 10 bits que se aplica como una VQ de división mediante dos tablas de códigos de cinco dimensiones, L2 y L3, de 32 entradas (5 bits cada una).

Con el fin de explicar el proceso de cuantificación, es conveniente primero describir el proceso de decodificación. Cada coeficiente se obtiene de la suma de dos tablas de códigos:

$$\hat{l}_i = \begin{cases} L1_i, (L1) + L2_i(L2) & i = 1, \dots, 5 \\ L1_i, (L1) + L3_{i-5}(L3) & i = 6, \dots, 10 \end{cases}, \qquad (6)$$

donde *L1*, *L2* y *L3* son índices de tabla de códigos. Para evitar resonancias bruscas en el filtro de síntesis cuantificado LP, los coeficientes $\hat{l}_i$ se ordenan de modo que los coeficientes adyacentes estén a una distancia mínima de *J*. La secuencia de reordenamiento se presenta a continuación:

$$\begin{aligned} &para\ i = 2, \dots, 10 \\ &\quad si\ (\hat{l}_{i-1} > \hat{l}_i - J) \\ &\quad\quad \hat{l}_{i-1} = \frac{(\hat{l}_i + \hat{l}_{i-1} - J)}{2} \\ &\quad\quad \hat{l}_i = \frac{(\hat{l}_i + \hat{l}_{i-1} + J)}{2} \\ &\quad fin \\ &fin \end{aligned}$$

Este proceso de reordenamiento se hace dos veces. La primera vez se establece un valor de *J* = 0,0012, y la segunda vez de *J* = 0,0006. Después de efectuado el reordenamiento, los coeficientes *LSF* cuantificados $\hat{\omega}_i^{(m)}$ para la trama presente *m* se obtienen de la suma ponderada de las salidas precedentes del cuantificador $\hat{l}_i^{m-k}$ y de la salida actual del cuantificador $\hat{l}_i^{m}$:

$$\hat{\omega}^{(m)} = \left(1 - \sum_{k=1}^{4} \hat{p}_{i,k}\right)\hat{l}_i^{m-k} + \sum_{k=1}^{4} \hat{p}_{i,k}\,\hat{l}_i^{m-k}, \qquad (7)$$

$$i = 1, \dots, 10,$$

donde $\hat{p}_{i,k}$ son los coeficientes del predictor MA conmutado. Para definir qué predictor MA ha de utilizarse se recurre a un bit, *L0* separado. En el

arranque, los valores iniciales de $\hat{l}_i^{(k)}$ están dados por $\hat{l}_i = i\pi/11$ para todo $k < 0$.

Una vez calculado $\widehat{\omega}_i$, se controla la estabilidad del filtro correspondiente. El procedimiento es el siguiente:

1. Se ordenan los valores del coeficiente $\widehat{\omega}_i$ de menor a mayor;
2. Si $\widehat{\omega}_i < 0.005$ entonces $\widehat{\omega}_i = 0.005$;
3. Si $\widehat{\omega}_{i+1} = \widehat{\omega}_i - 0.0391 < 0.005$, entonces $\widehat{\omega}_{i+1} = \widehat{\omega}_i + 0.0391$, siendo $i = 1, \ldots, 9$;
4. Si $\widehat{\omega}_{10} > 3.135$ entonces $\widehat{\omega}_{10} = 3.135$.

El procedimiento para codificar los parámetros *LSF* puede resumirse como sigue. Se procurará hallar para cada uno de los dos predictores MA la mejor aproximación a los coeficientes *LSF* del momento. Como mejor aproximación se considera la que minimiza el error cuadrático medio ponderado:

$$E_{lsf} = \sum_{i=1}^{10} \omega_i (\omega_i - \widehat{\omega}_i)^2. \quad (8)$$

Los pesos $\widehat{\omega}_i$ se hacen adaptativos como función de los coeficientes LSF no cuantificados:

$$\omega_i = \begin{cases} 1{,}0 \; Si \; (\omega_2 - 0.04\pi - 1) > 0 \\ 10(\omega_2 - 0.04\pi - 1)^2 En \; los \; demás \; casos \end{cases}'$$

$$\omega_i = \begin{cases} 1{,}0 \; Si \; (\omega_{i+1} - \omega_{i-1} - 1) > 0 \\ 10(\omega_{i+1} - \omega_{i-1} - 1)^2 En \; los \; demás \; casos \end{cases}, \quad (9)$$

para $2 \leq i \leq 9$

$$\omega_{10} = \begin{cases} 1{,}0 \; Si \; (-\omega_9 + 0.92\pi - 1) > 0 \\ 10(\omega_9 - 0.92\pi - 1)^2 En \; los \; demás \; casos. \end{cases}$$

Además, los pesos $\widehat{\omega}_5$ y $\widehat{\omega}_6$ se multiplican cada uno por 1,2.

El vector a cuantificar para la trama actual *m* se obtiene mediante:

$$l_i = \frac{\left[\omega_i^{(m)} - \sum_{k=1}^{4} \hat{p}_{i,k} l_i^{(m-k)}\right]}{\left[1 - \sum_{k=1}^{4} \hat{p}_{i,k}\right]} \; i = 1, \ldots, 10. \quad (10)$$

Se indaga la primera tabla de códigos L1, seleccionando la entrada *L*1 que hace mínimo el error cuadrático medio (no ponderado). Seguidamente se indaga la segunda tabla de códigos, L2, que define la parte inferior de la segunda fase. Para cada posible candidato, se reconstruye el vector parcial $\widehat{\omega}_i, \; i = 1, \ldots, 5$ mediante la ecuación (7), reordenándolo para asegurar una distancia mínima de 0,0012. Se calcula el error cuadrático medio ponderado según la ecuación (8) y se selecciona el vector *L*2 que produzca el error más bajo. Aplicando el vector *L*1 seleccionado de la primera fase y la parte inferior del *L*2 de la segunda fase, se extrae la parte superior de la segunda fase de la tabla de códigos L3. Se procede a reordenar nuevamente los valores para asegurar una distancia mínima de 0,0012. Se selecciona el vector *L*3 que minimice el error cuadrático medio ponderado. El vector resultante $\hat{l}_i, \; i = 1, \ldots, 10$ se reordena para asegurar una distancia mínima de 0,0006. Este proceso se lleva a cabo para cada uno de los dos predictores de media móvil, definidos por *L0*, seleccionando el predictor de MA *L*0 que produzca el menor error cuadrático medio ponderado. Como se indica al comienzo de este epígrafe, el reordenamiento del vector resultante $\hat{l}_i$ se efectúa dos veces, procediéndose a un control de estabilidad para obtener los coeficientes *LSF* cuantificados $\widehat{\omega}_i$.

A partir de estos vectores: a_p y $\widehat{\omega}_i$ se realizará la función de reconocimiento, o sea, se convierten en parámetros de decisión.

De esta manera será posible evaluar la factibilidad de realizar el control remoto por voz sobre parámetros cuantificados o sobre parámetros que no han sido previamente cuantificados. Esto es importante ya que el éxito del reconocimiento depende en gran medida de la información espectral de la señal de voz que se pueda recuperar de los parámetros de compresión.

Otro aspecto de interés será evaluar la cantidad de coeficientes de predicción lineal que optimiza el proceso de reconocimiento, sin llegar

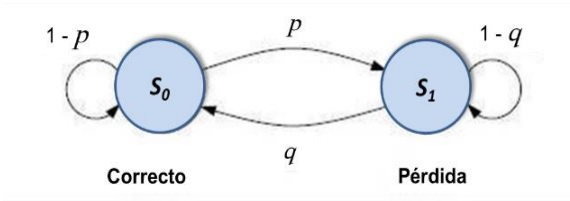

Fig. 2. Modelo de Gilbert–Eliot

a afectar el proceso por un exceso de requerimientos computacionales.

– Simulación cómo método:

Como método práctico general se ha empleado el método de la simulación, por cuanto ha sido necesario:

1. El empleo de los codificadores empleados para la compresión en ambientes virtuales.
2. Simular el comportamiento de las pérdidas de información en el canal de transmisión, que como se ha dicho, es un canal IP en el que se considerará que los paquetes se pierden en ráfaga con una probabilidad de pérdida total conocida.
3. La validación del comportamiento de los parámetros del proceso de reconocimiento automático del habla.

Todo lo anteriormente mencionado se realizó empleando el asistente matemático MATLAB R2014a, que es un lenguaje de alto nivel con ambiente interactivo para: computación numérica, visualización y programación.

– Simulación de pérdidas en el canal de transmisión:

Para lograr esta simulación de pérdidas de información en el canal se emplea el modelo de Gilbert–Eliot de dos estados, tal como se muestra en la Figura 2. A diferencia del modelo de Gilbert, este introduce un nuevo parámetro pel como probabilidad de pérdida total, permitiendo ajustar los patrones de pérdidas a las trazas reales, donde p y q son las probabilidades de pasar del estado s_0 al estado s_1 y viceversa respectivamente. Mientras que $1-p$ y $1-q$ son las probabilidades de transición de un estado a otro cuando s_0 y s_1 son los estados actuales en cada caso [7].

Para este caso las probabilidades de recibir, P_l y perder $\overline{P}_l$, una ráfaga de l paquetes consecutivos vienen determinadas por:

$$P_l = p(1-p)^{l-1}, \tag{11}$$

$$\overline{P}_l = q(1-q)^{l-1}. \tag{12}$$

y la duración media de las ráfagas de pérdidas es [8]:

$$L_{ráfaga} = \sum_{l=1}^{\infty} l * \overline{P}_l = \sum_{l=1}^{\infty} l * q * (1-q)^{l-1} = \frac{1}{q}. \tag{13}$$

– Tipos de Reconocimiento Automático del Habla (RAH):

En resumen, el RAH puede clasificarse considerando varios criterios, en opinión de los autores de este trabajo los más relevantes son:

– Dependencia del locutor: puede ser dependiente del locutor, o independiente de este. O sea, en el primer caso el sistema puede tratar voz perteneciente a un conjunto cerrado de locutores, y en el otro no, sin embargo el conjunto de locutores del entrenamiento debe ser amplio “de modo que el reconocimiento de un nuevo locutor sea suficientemente preciso” [9,10].
– Atendiendo a los tipos de locuciones empleadas: reconocedores de palabras aisladas y de voz continua.

Por otra parte, desde el punto de vista de las funciones que realizan los componentes del sistema, pueden implementarse distintas arquitecturas [11]:

– Sistema de reconocimiento empotrado, (ESR, *Embedded Speech Recognition*), integra el sistema de reconocimiento al completo en el cliente.
– Arquitectura sólo servidor, (NSR, *Network-Based Speech Recognition*), solo lleva a cabo el proceso de adquisición en el terminal cliente [12].
– Arquitectura cliente-servidor (DSR, *Distributed Speech Recognition*), las tareas de procesamiento y cómputo se encuentran

distribuidas entre el terminal y el servidor remoto de reconocimiento [13].

Tomando en cuenta estas clasificaciones, esta investigación se enfoca en reconocimiento de palabras aisladas y dependiente del locutor en sistemas DSR, ya que las unidades básicas a reconocer serán palabras, las características espectrales captadas por los coeficientes $\propto_k$ (cuantificados o no) permiten una mayor precisión para hablantes individuales, y las funciones de reconocimiento se encontrarán distribuidas entre cliente y servidor.

– El método de decisión:

La esencia del reconocimiento se basa en una comparación. Se empleará la distancia euclidiana, como un tipo de distancia "ordinaria" entre dos puntos de un espacio euclídeo[1], la cual se deduce a partir del teorema de Pitágoras, según (14):

$$d(A,B) = \sqrt{\sum_{i=1}^{N}(b_i - a_i)^2} = \sqrt{(b_1 - a_1)^2 - (b_2 - a_2)^2 + \cdots + (b_N - a_N)^2}. \quad (14)$$

La información de una unidad básica de reconocimiento se comparará con una referencia obtenida en una etapa de entrenamiento previa. En este caso, ambas serán vectores cuyos elementos son los coeficientes de predicción lineal.

4. Planteamiento del problema

La cuestión a resolver con este trabajo responde a: ¿cómo aprovechar las bondades de la predicción lineal para realizar un reconocimiento del habla?

De forma general, el objetivo que persigue este trabajo es: determinar las condiciones bajo las cuales resulta factible realizar un reconocimiento automático del habla de forma remota, cuando la voz ha sido comprimida y transmitida por canales basados en el Protocolo de Internet y caracterizados por la ocurrencia de pérdidas de información.

De forma particular, se pretende valorar la factibilidad de realizar dicho control cuando:

– Los parámetros de compresión, dígase los coeficientes $\propto_k$, se encuentran cuantificados o no.
– Se aumenta la cantidad de coeficientes de predicción.
– Las pérdidas en el canal de transmisión varían su comportamiento.

Además se tendrá en cuenta el tiempo promedio de reconocimiento para cada condición experimental.

Para ello se tendrán dos casos de estudio: el códec LPC-10 y el G.729, que permitirán evaluar las condiciones anteriormente mencionadas. Para lograrlo no será necesario emplear la totalidad de las instrucciones de los correspondientes algoritmos de dichos códecs, solamente una porción de la parte codificadora, hasta la determinación de los parámetros $\propto_k$ y $\widehat{\omega}_i$, y en el caso del códec G.729, será necesaria la recuperación de los parámetros cuantificados.

5. Discusión de los resultados

Se diseñó una etapa de entrenamiento que consistió en generar una base de datos con la información de reconocimiento extraída a un conjunto de cien palabas en idioma español, las cuales fueron escogidas de forma aleatoria de una muestra de 1000 palabras entre: agudas, graves, esdrújulas y sobre-esdrújulas.

La base de datos consistía en la palabra y sus respectivos coeficientes $\propto_k$ y $\widehat{\omega}_i$ según sea el caso experimental (I) o (II) descritos en la sesión anterior.

Las señales vocales fueron adquiridas con los siguientes parámetros: una frecuencia de muestreo: 8KHz, 256 niveles y tomando un solo canal.

Aunque el tiempo de duración asignado a las palabras grabadas en el entrenamiento fue de un segundo, lo cual devuelve un vector de 500 $\propto_k$ de longitud para el LPC-10 y de 1000 $\propto_k$ para el G.729 (que serán convertidos a $\widehat{\omega}_i$), para realizar el reconocimiento fue necesario implementar una

[1] Un tipo de espacio geométrico donde se satisfacen los axiomas de Euclides de la geometría.

alineación temporal dinámica, con el objetivo de *"medir"* la similitud de: la información a reconocer (*A*) con respecto a la información en la base de datos (*B*).

Para encontrar la función de alineamiento a partir de los vectores A y B, siendo:

$$A = \{a_1, a_2, \dots, a_i, \dots, a_M\}, \quad (15)$$

$$B = \{b_1, b_2, \dots, b_j, \dots, b_N\}. \quad (16)$$

Llamando C a la función de alineamiento:

$$\mathrm{C} = \{\mathrm{c}(1), \mathrm{c}(2), \dots, \mathrm{c}(\mathrm{k}), \dots, \mathrm{c}(\mathrm{k})\}, \quad (17)$$

donde $\mathrm{c}(\mathrm{k})$, es un par de punteros a los elementos a comparar:

$$\mathrm{c}(\mathrm{k}) = [\mathrm{i}(\mathrm{k}), \mathrm{j}(\mathrm{k})]. \quad (18)$$

Para cada $\mathrm{c}(\mathrm{k})$ se tiene una función de coste:

$$\mathrm{d}\{\mathrm{c}(\mathrm{k})\} = \delta\left(\mathrm{a}_{\mathrm{i}(\mathrm{k})}, \mathrm{b}_{\mathrm{i}(\mathrm{k})}\right). \quad (19)$$

Esta función refleja la discrepancia entre los elementos comparados. En este trabajo, como se mencionó anteriormente, la función de coste empleada es la distancia euclidiana:

$$\mathrm{d}\{\mathrm{c}(\mathrm{k})\} = \sqrt{\left(\mathrm{a}_{\mathrm{i}(\mathrm{k})} - \mathrm{b}_{\mathrm{i}(\mathrm{k})}\right)^2}. \quad (20)$$

Siendo, finalmente, la función de alineamiento, aquella que minimice la función de coste total:

$$D(C) = \sum_{k=1}^{K} d\{c(k)\}. \quad (21)$$

Este procedimiento fue implementado en la etapa de reconocimiento, para permitir una mayor efectividad aun cuando un mismo locutor emplee diferentes intervalos de tiempo para pronunciar una misma palabra.

Tabla 1. Efectividad del reconocimiento por tipo de palabra pronunciada.

Tipo de Palabra Pronunciada	Efectividad de reconocimiento (%)
Agudas	99%
Graves	97%
Esdrújulas	95%
Sobre – esdrújulas	93%

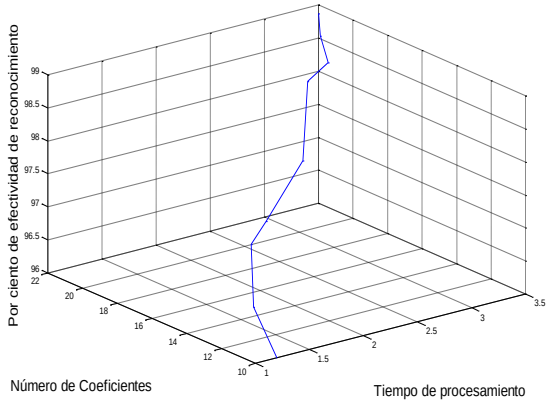

Fig. 3. Relación del por ciento de efectividad total con respecto al coste computacional y al número de coeficientes

Además se realizaron variaciones del número de coeficientes $\propto_k$ determinados para cada trama de análisis en cada uno de los códec.

5.1. Resultados LPC-10

Para el caso experimental (I), sin pérdidas en el canal, las pruebas realizadas arrojaron los siguientes resultados.

La efectividad del reconocimiento por tipo de palabra, se muestra en la Tabla 1, para un promedio del 96%.

Sin embargo, al aumentar el número de coeficientes extraídos a cada segmento de 20mseg de voz, el comportamiento de la efectividad de reconocimiento, teniendo como referencia además el coste computacional, se comportó tal como se muestra en la Figura 3.

El porcentaje de efectividad total aumenta con respecto al tiempo de procesamiento, a medida que se aumenta el número de coeficientes

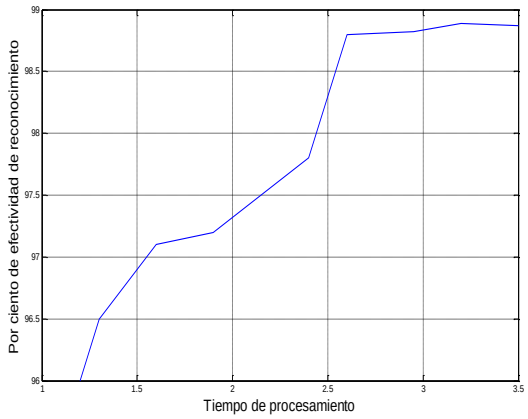

Fig. 4a. Relación del % de efectividad vs. Tiempo de procesamiento

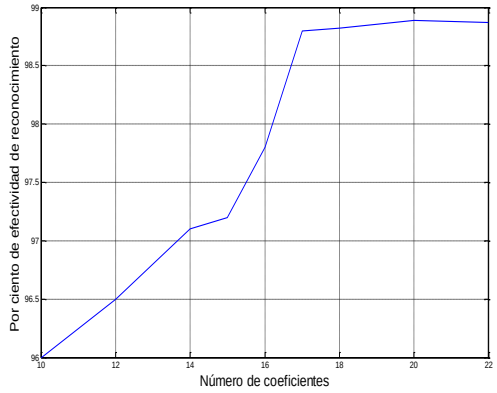

Fig. 4b. Relación del % de efectividad vs. Número de coeficientes determinados

calculados. Sin embargo, cuando se alcanza la cantidad de 17 coeficientes, este parámetro se mantiene constante.

Tal como se ilustra en la Figura 4, a partir de los 17 $\propto_k$ el reconocimiento sigue teniendo la misma efectividad, la cual es muy buena ya que se acerca al 99%. Lógicamente, esto tiene como consecuencia un aumento del costo computacional en términos de tiempo de procesamiento.

La figura siguiente confirma esta situación, ahora en relación al número de coeficientes determinados.

La efectividad de reconocimiento aumenta considerablemente hasta 17 coeficientes, pero una vez que alcanza esta cantidad permanece constante alrededor del 98.8%.

Para estudiar más a fondo el comportamiento del sistema se consideró la ocurrencia de pérdidas en el canal de transmisión. Para cada valor de probabilidad de pérdida se realizaron una serie de pruebas, en las cuales se aumentó el número de coeficientes determinado. Los resultados obtenidos son mostrados en la gráfica de la Figura 5, para valores de probabilidad de pérdidas entre el 10 y el 30%, y para cada uno de ellos se emplearon entre 10 y 22 coeficientes.

5.2. Resultados G.729

Para el caso experimental (II), con el codificador G.729, y bajo condiciones ideales de canal de transmisión, los resultados fueron muy similares al caso experimental (I), aunque ligeramente por encima. Esto se justifica por el hecho que este códec trabaja con segmentos de la señal de voz que representan la mitad de la duración de aquellos con que trabaja el LPC-10, o sea, 10mseg, por tanto, en la unidad de tiempo (1 segundo) se realizan el proceso de extracción de características el doble de la cantidad de veces.

La diferencia más significativa estuvo en cuanto al coste computacional, ya que los procesos de cuantificación y recuperación de los coeficientes, consumen más tiempo de procesamiento.

La figura siguiente muestra una comparación de los resultados obtenidos en ambos casos experimentales, para todas las pruebas realizadas.

Cabe destacar que le efectividad de reconocimiento de las palabras tiene una efectividad satisfactoria, incluso cuando se pierde una parte considerable del volumen total de la información.

En cuanto al tiempo de procesamiento, es evidente que el coste que implica usar el G.729 es mayor que al emplear el LPC-10, y aumenta cuando se incrementa el número de coeficientes calculados.

Todo esto permite afirmar que el empleo de la variante del LPC-10 es una propuesta robusta para el reconocimiento de palabras.

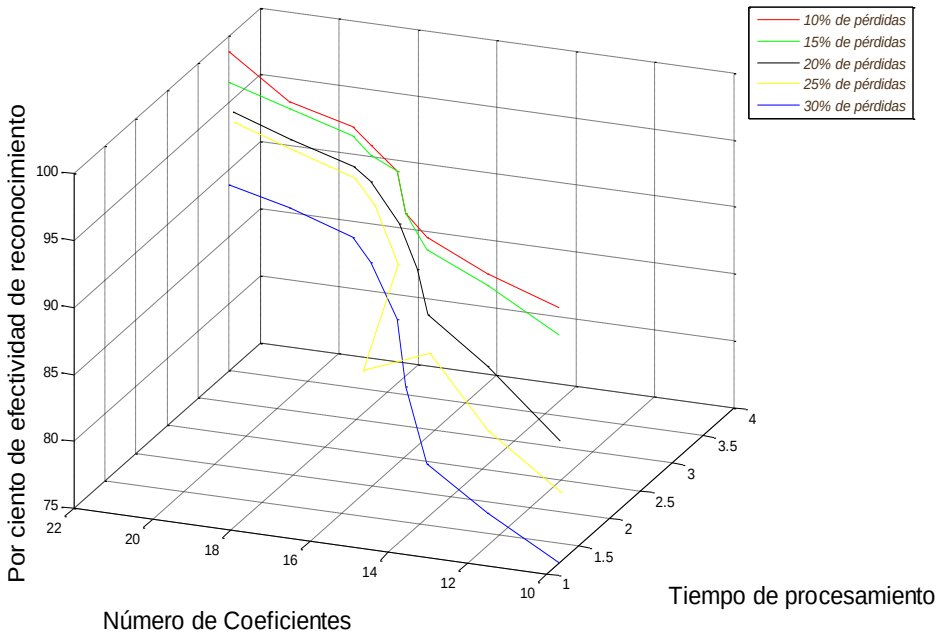

Fig. 5. Resultados del caso experimental (I) al aumentar el por ciento de probabilidad de pérdidas en el canal

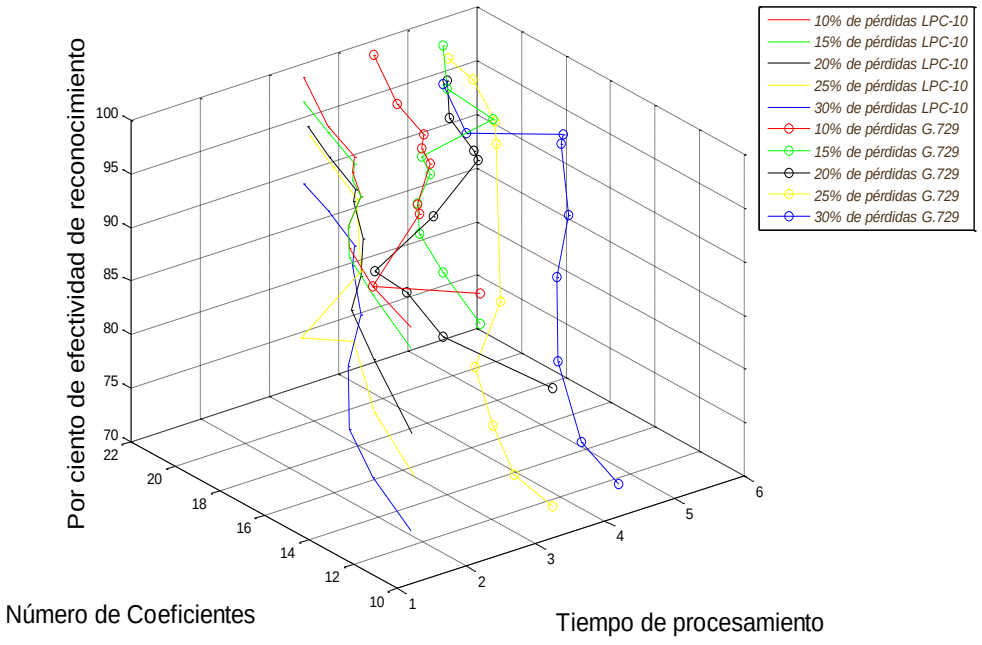

Fig. 6. Resultados de ambos casos experimentales

6. Trabajos futuros

Como perspectiva de aplicación de estos resultados en futuras investigaciones, está en primer lugar, la implementación de los métodos propuestos en esquemas de reconocimiento continuo del habla.

De modo que el reconocimiento de la palabra como unidad fonética básica y empleando predicción lineal, sea aplicable al reconocimiento del discurso continuo.

Además, los propios resultados obtenidos, pueden ser perfectibles a emplear métodos más precisos de decisión, o incluso, el empleo de un nuevo parámetro que caracterice la unidad fonética empleada.

7. Conclusiones

La predicción lineal se muestra como un método efectivo para realizar el reconocimiento de palabras aisladas.

En este reconocimiento el proceso de cuantificación de parámetros solo implica un aumento considerable del coste computacional.
Las pérdidas de información que puedan ocurrir en esquemas de reconcomiendo distribuido, no afectan la efectividad del reconocimiento al emplear la predicción lineal, por lo que esta técnica se presenta como un método robusto para el reconocimiento automático del habla.

Agradecimientos

Este trabajo ha sido desarrollado con la colaboración y apoyo del Departamento de Telecomunicaciones y Electrónica de la Universidad de Pinar del Río Hermanos Saíz, así como de la administración de la Facultad de Ciencias Técnicas de la propia institución. Los autores reconocen y agradecen los comentarios y sugerencias de los revisores.

Referencias

1. **Rabiner, L. & Juang, B. (1993).** *Fundamentals of speech recognition*. Prentice-Hall International.
2. **Deller, J.R., Proakis, J.G., & Hansen, J.H.L.** *Discrete-time processing of speech signals*. Prentice-Hall.
3. **Sahab, A.R. & Khosroo, M. (2008).** Speech Coding Algorithms: LPC10, ADPCM, CELP and VSELP. *Journal of Applied Mathematics,* Islamic Azad University of Lahijan, Vol. 5, No. 16.
4. **Telecommunication Standardization Sector of ITU. (2007).** *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP).* ITU-T Recommendation G.729.
5. **Tremain, T.E. (1989).** The government standard linear predictive coding algorithm: Lpc-10. *Speech Technology*.
6. **Proakis, J.G. & Manolakis, D.G. (2001).** *Digital Signal Processing: Principles, Algorithms and Applications*. Prentice-Hall.
7. **Docío, L., Cardenal, A., & García, C. (2004).** *Speech Recognition over Digital Channels: Robustness and Standards*. John Wiley and Sons.
8. **Jiang & Schulzrinne, H. (2000).** Modeling of packet loss and delay and their elect on Real-Time multimedia service quality. *Proceedings of NOSSDAV*.
9. **Carmona, J.L. (2009).** *Reconocimiento de voz codificada sobre redes IP*. Editorial de la Universidad de Granada.
10. **Rabiner, L. & Juang, B.H. (1998).** *Digital Signal Processing Handbook*. CRC Press.
11. **Tan, Z. & Varga**, **I. (2008).** Automatic Speech Recognition on Mobile Devices and over Communication Networks. Chapter in *Network, Distributed and Embedded Speech Recognition: An Overview*, Springer-Verlag.
12. **Peinado, A.M. & Segura J.C. (2006).** *Speech Recognition over Digital Channels: Robustness and Standards*. John Wiley and Sons.
13. **Haavisto, P. (2006).** Speech recognition for mobile communications. *Proceedings of the COST Workshop on Robust Methods for Speech Recognition in Adverse Conditions,* IEEE Speech Technical Committee Newsletter.

Article received on 09/10/2015; accepted 20/05/2016.
Corresponding author is Elieser E. Gallego.